Psychological Test and Assessment Modeling, Volume 53, 2011 (4), 414-439

Item banking for C-tests: A polytomous Rasch modeling approach

Thomas Eckes[1]

Abstract

C-tests are gap-filling tests widely used to assess general language proficiency for purposes of placement, screening, or provision of feedback to language learners. C-tests consist of several short texts in which parts of words are missing. Development, administration, and scoring of C-tests are particularly efficient when use is made of a calibrated item bank. Rasch measurement provides a powerful approach to item bank construction. Based on construing C-test texts as superitems, where item values correspond to the number of gaps in the text filled in correctly, two polytomous Rasch models were applied to analyze and evaluate a large set of texts: Andrich's (1978) rating scale model and Müller's (1987, 1999b) continuous rating scale model. The test construction phase comprised a total of 218 texts trialled in 27 independent samples, covering a total of 5,927 participants. Across samples, reliability indices ranged from .94 to .98. Texts showing unsatisfactory model fit or DIF were eliminated. The remaining texts were put on the same difficulty scale through a concurrent estimation procedure. Results clearly attested to the suitability of polytomous Rasch models to calibrate texts for purposes of item banking.

Key words: polytomous Rasch models, item banking, language testing, C-test, local item dependence

[1] *Correspondence concerning this article should be addressed to:* Thomas Eckes, PhD, TestDaF Institute, Massenbergstr. 13 b, 44787 Bochum, Germany; email: thomas.eckes@testdaf.de

C-tests are written tests of general language proficiency (Grotjahn, Klein-Braley, & Raatz, 2002; Klein-Braley, 1997). As a rule, C-tests consist of four to eight short authentic texts in which parts of words are missing. Examinees have to insert the missing parts, that is, to restore the original words in each text. C-tests have frequently been used in educational and occupational contexts, mainly serving purposes of placement, screening, or providing feedback to language learners regarding their current level of proficiency (see, e.g., Harsch & Schröder, 2007; Norris, 2006; Reichert, Keller, & Martin, 2010).

Like the classic cloze test from which they were developed, C-tests build on the principle of reduced redundancy (Klein-Braley, 1997; Sigott, 2004; Spolsky, 1971): The redundancy inherent in natural language is deliberately reduced by deleting parts of words in a text according to a well-defined rule. When examinees fill in the gaps, they have to draw on their knowledge of the target language, in particular on their lexical and grammatical knowledge. Accordingly, the more differentiated, comprehensive, and accessible this knowledge is, the better the examinees will perform on the test.

Research into the psychometric quality of C-tests, mostly with English as the target language, has provided ample evidence of high test reliability and validity (Eckes & Grotjahn, 2006a; Sigott, 2004). For example, C-tests have been shown to correlate significantly both with receptive skills (reading, listening) and with productive skills (writing, speaking), most of these correlations ranging from .50 to .70. In a series of confirmatory factor analyses, the pattern of correlations between a German C-test and the four language skills was best accounted for by a single factor representing general language proficiency (Eckes & Grotjahn, 2006a).

Building on this line of research, the present paper focuses on a Rasch measurement approach to constructing a calibrated item bank for C-tests. In a calibrated item bank the parameter estimates for all items in the bank have been placed on the same difficulty scale (see, e.g., Szabó, 2008; Vale, 2006; Wright & Stone, 1999). Calibrated item banks have optimal functionality in terms of (a) efficiency of access to items and their psychometric attributes, (b) flexibility of test development, and (c) ease of test administration and scoring, including the implementation of web-based testing.

When constructing a calibrated item bank for C-tests, an important issue concerns the choice of a suitable psychometric model for item calibration and linking. It is suggested here that polytomous Rasch models (Embretson & Reise, 2000; Ostini & Nering, 2006; Wright & Masters, 1982) provide a particularly promising approach. Drawing on a large set of C-test data gathered in the context of developing an online placement test of German as a foreign language, the use of such models for item-banking purposes is investigated.

Polytomous Rasch models for C-tests

Local item dependence

Due to the text-based design of C-test construction, gaps within a text are directly or indirectly linked to each other by content and various linguistic attributes. Therefore, a considerable degree of dependence typically exists between the gaps. A major part of this dependence is captured by the term *passage dependence* (Yen, 1993). This means that several items (gaps) are attached to the same passage, and information that is used to respond to each of these items is interrelated in the passage. Another cause of dependence is *item chaining* (Yen, 1993); that is, items (gaps) are organized in a stepwise fashion such that knowing the answer to one item increases the chances of knowing the answer to the next item.

The dependence inherent in a C-test violates a basic assumption underlying the application of standard psychometric models. This is the assumption of local (or conditional) independence, saying that for persons at the same level of ability in the variable being measured, responses to any given item are independent of responses to other items on the test (Henning, 1989; Lord & Novick, 1968). An analysis that ignores possible local item dependence (LID) runs the risk of overestimating the precision of examinee proficiency measures and may yield biased item difficulty parameters (E. V. Smith, 2005; Yen & Fitzpatrick, 2006).

In order to deal with the LID problem related to C-tests, Raatz (1985) and Grotjahn (1987) suggested to consider each text as a *superitem* and to compute the reliability of the test on the basis of superitems only. The term *superitem* was first employed by Cureton (1965) to denote subsets of items within a test, that is, items that share a common stimulus or are linked by common content. Much the same meaning is conveyed by the term *testlet* (Wainer & Kiely, 1987) introduced in the context of computerized adaptive testing. Wainer and Kiely construed testlets as aggregations of items, which by their very nature are not conditionally independent. In a similar vein, Rosenbaum (1988) proposed to formulate the independence requirement between *item bundles* rather than between individual items (see also Wilson & Adams, 1995).

Building on the notion of texts as superitems (or testlets or item bundles) the question arises as to which kind of measurement model would be suited to provide the desired item calibrations. In the next section, two polytomous Rasch models that have attracted some attention in the field are briefly discussed (Baghaei, 2008, 2011; Eckes, 2006, 2007, 2010a, 2010b; Eckes & Grotjahn, 2006b; Lee-Ellis, 2009). The first model belongs to the class of discrete response models: the *rating scale model* (Andrich, 1978); the second one belongs to the class of continuous response models: the *continuous rating scale model* (Müller, 1987, 1999b).

Discrete response models

When a C-test text is construed as a superitem, item values correspond to the number of gaps filled in correctly. For example, a text that contains 20 gaps would be considered as one large item with item values ranging from 0 to 20. More generally, a text that contains m gaps could take on $m + 1$ item values. Hence, a C-test text can also be construed as a polytomous item or, alternatively, as a (discrete) rating scale with $m + 1$ successive categories k (i.e., $k = 0, 1, \dots, m$).

The two most widely used Rasch models for discrete polytomous response data are Andrich's (1978) rating scale model (RSM) and Masters' (1982) partial credit model (PCM). Both models add parameters to the basic Rasch model for dichotomous data (Rasch, 1960/1980) that describe the functioning of the rating scale. The RSM adds a threshold parameter to represent the relative difficulty of the transition between adjacent response categories. That is, the threshold parameters are located at the intersections of the probability curve of one category with the probability curve of the next. Because the RSM employs only one set of threshold parameters across all items on a test, its use presupposes that items have a common rating scale structure; that is, all items have the same number of response categories, and the relative difficulty between categories is constant.

When items have differing numbers of response categories or when the relative difficulty between categories is expected to vary from item to item, the PCM may be considered a suitable alternative. The PCM estimates threshold parameters for each item separately, allowing each item to have a unique rating scale structure.

However, when it comes to calibrating C-test texts, the estimation of additional item-specific sets of threshold parameters sets sample size requirements that are hard to meet in practice. Following Linacre (2004), there should be at least 10 observations of each response category. Depending on (a) the number of texts, (b) the number of gaps per text, and (c) the difficulty of texts relative to the proficiency of examinees, the minimum sample size needed for a PCM analysis of a single set of C-test data may easily exceed 500 persons (see also Embretson & Reise 2000). Collapsing less frequently observed response categories may be a meaningful option only when the distribution of person parameters and the distribution of item parameters are closely aligned to each other.

In each of the samples examined in the present research, sample size was much too small to be considered sufficient for a PCM analysis. Therefore, building on previous research on the Rasch analysis of C-tests (see Eckes, 2006, 2010b), it was decided to employ the RSM. Note also that the RSM has recently been shown to be highly robust against violations of the constant threshold assumption (Baghaei, 2010).

In the RSM, the probability that person n with ability θ_n will obtain a score of k (k = 0, … , m) on item i is expressed as

$$p_{ik}(\theta_n) = p(x_{in} = k \mid \theta_n) = \frac{\exp \sum_{j=0}^{k} \left[\theta_n - (\beta_i + \tau_j) \right]}{\sum_{r=0}^{m} \exp \sum_{j=0}^{r} \left[\theta_n - (\beta_i + \tau_j) \right]}, \tag{1}$$

where $\tau_0 \equiv 0$ (Wright & Masters, 1982).

In Equation 1, β_i represents the difficulty parameter for item (or superitem) i, and τ_j represents the threshold parameter for category j, that is, the parameter for the transition from category j –1 to category j. Item i has m categories, and k is the count of the number of successfully completed categories for that item. That is, in the present context, k is the count of gaps within text i that person n filled in correctly.

Continuous response models

Construing texts within a C-test as polytomous items or rating scales implies that the number of possible response categories is fairly large, much larger than the usually recommended four to seven categories (see, e.g., Lozano, García-Cueto, & Muñiz, 2008; Preston & Colman, 2000). When the number of categories becomes infinitely large, continuous rating scales result as a limiting case. Such scales are also known as graphic rating scales or visual analogue scales; in the context of Internet-delivered questionnaires, continuous rating scales appear under such names as “sliders” or “slider bars”.

For psychometric modeling purposes, gap-filling texts may be viewed as approximating continuous rating scales and thus be analyzed by means of a Rasch model for continuous ratings. One such model is Müller’s (1987, 1999b) continuous rating scale model (CRSM). This model is a direct extension of Andrich’s (1978) RSM.

The CRSM assumes a response mechanism where a latent response variable, originally unbounded and following a normal distribution, is doubly truncated to fit the response format constraint. Specifically, Müller (1987) considered the rating scale as a straight line segment of midpoint *c* and length *d*. In keeping with Samejima (1973), the end points of that scale, that is, $c \pm d/2$, are assumed to be defined (e.g., by labels such as “extremely positive” and “extremely negative”), and the person is allowed to mark any point along the line segment. Given this, the CRSM is defined as follows (Müller, 1987, 1999b):

Let the number of categories of the rating scale grow to infinity (i.e., $m \rightarrow \infty$). Then, the result is a continuous random variable X that can take on any value $x \in [c - d/2, c + d/2]$, and the probability that person *n* with ability θ_n will show a response on item *i* within a given interval $[a, b]$ of the line segment is given by

$$p(a \le X_{ni} \le b | \theta_n, \beta_i, \lambda) = \frac{1}{\gamma(\theta_n, \beta_i, \lambda)} \int_a^b \exp\left[x_{ni}(\theta_n - \beta_i) + x_{ni}(2c - x_{ni})\lambda\right] dx_{ni}, \qquad (2)$$

where

$$\gamma(\theta_n, \beta_i, \lambda) = \int_{c-d/2}^{c+d/2} \exp\left[t(\theta_n - \beta_i) + t(2c - t)\lambda\right] dt. \qquad (3)$$

The probability density of the continuous response variable is

$$f(x_{ni} | \theta_n, \beta_i, \lambda) = \frac{1}{\gamma(\theta_n, \beta_i, \lambda)} \exp\left[x_{ni}(\theta_n - \beta_i) + x_{ni}(2c - x_{ni})\lambda\right]. \qquad (4)$$

In Equations 2 to 4, the λ term represents the dispersion parameter. The CRSM posits a uniform density of thresholds along the latent interval $[-\lambda d, \lambda d]$. Hence, λ parameterizes the range of the threshold distribution.

The dispersion parameter indicates the degree to which the thresholds increase in a strictly monotonic fashion along the continuous rating scale. If respondents actually use the continuous scale in a continuous manner, the dispersion parameter takes on positive values greater than zero (i.e., $\lambda > 0$; the “regular case” of the model). Conversely, violations of model assumptions are indicated by negative values of the dispersion parameter (i.e., $\lambda < 0$; the “irregular case”). In the special case of $\lambda = 0$ the uniform threshold distribution degenerates into a single threshold $\tau = 0$ (the “degenerate case”; Müller, 1987, 1999b).

A different approach was proposed by Linacre (2001). This approach rests on a highly general continuous Rasch model, in which the percentage scale, with the range of 0% to 100%, reported with integers, is considered as approximating a continuous rating scale with 101 categories. In that model, different forms of the measurement function are possible. Particular forms may include polynomials, trigonometric, and logarithmic functions. To my knowledge, none of these functions have yet been employed in research or applied settings.

Overview of the present research

As part of an ongoing process of developing a calibrated item bank for use with an online placement test of German as a foreign language, data from 27 independent samples of participants were analyzed. In each sample, participants worked on a different set containing 10 texts with 20 gaps each. In order to evaluate texts within a given set and to jointly calibrate texts across all sets, a two-stage procedure was employed. In the first stage, the data within each sample were analyzed separately by means of two polytomous Rasch models, that is, Andrich’s (1978) RSM and Müller’s (1987) CRSM. Based on statistical indicators of model fit, texts that did not function properly were rejected. In the second stage, all texts that came through the first stage were put on the same difficulty scale using a concurrent estimation procedure. The resulting difficulty estimates provided the key input data for the item bank.

Method

Participants

A total of 5,927 participants volunteered to work on sets of mutilated German texts. There were 3,792 females and 2,093 males; 42 participants did not indicate their gender. The age of 83.4% of the total sample of participants ranged from 18 to 28 years (M = 23.30, SD = 6.05), 4.1% of the participants were younger than 18 years, 12.4% were older than 28 years.

At the time of testing, participants were either attending German language courses as part of a preparatory study program in Germany or planning to study at a German university while still in their home country. Texts were administered at test centers of the TestDaF Institute (www.testdaf.de) or at lectorates of the German Academic Exchange Service (www.daad.de) in 46 countries from around the world.

Participants came from 125 different countries. In terms of the number of participants, the following ten national groups ranked highest (percentage in parentheses): Russia (11.4%), Indonesia (7.9%), People's Republic of China (5.7%), Lithuania (4.8%), Poland (4.5%), Bulgaria (3.9%), Ukraine (3.9%), Morocco (3.0%), Turkey (2.9%), and France (2.8%).

Following data analysis, each participant received feedback on his or her performance. Feedback consisted of the test score earned in a given set of texts and the percentage rank achieved in the sample the participant belonged to.

Test material

A total of 218 mutilated texts, each containing 20 gaps, were subjected to detailed examination in a series of trial studies spanning a three-year period. Texts were compiled in sets of 10 texts, making a total of 27 sets. Within each set, texts were arranged in ascending order of supposed difficulty based on findings from pre-testing and expert judgment. The main purpose of this tentative difficulty-based ordering of texts was to keep low-proficient participants from becoming deterred when having to start with a text that was unduly hard. As usual, each text within a given set dealt with a different topic.

Texts were constructed according to the classic deletion rule (the "rule of two"); that is, words were mutilated by deleting the second half of every second word, beginning with the second word of the second sentence. If a word had an odd number of letters, the larger part was deleted (see Grotjahn et al., 2002). Throughout the texts, the missing part of each word was indicated by a single underline of constant length. The instruction read: "Complete the gaps in the following texts in a meaningful way. You have five minutes for each text". The time allowed was also printed above each text. Test administrators strictly controlled adherence to this time limit.

The Appendix presents two texts illustrating the exact C-test format used in the present study. Note that these texts have been taken from a sample test included in the online placement test's website for practice purposes (see www.ondaf.de, link "Beispieltest"). Moreover, both texts had been calibrated following the same psychometric principles as discussed in this report. In that calibration, Sample Text 1 proved to be fairly easy, difficulty estimate = –1.39 logits (*SE* = 0.04); Sample Text 2 was much more difficult (for learners of German as a foreign language), difficulty estimate = –0.48 logits (*SE* = 0.04).

Across all trial sets, two texts were the same. These common texts served to provide the link between the different sets, following a non-equivalent groups with anchor test design (also called common-item non-equivalent groups design; see, e.g., Kolen & Brennan, 2004; Wolfe, 2000; Wright & Stone, 1979). In each set, the common texts appeared at the third and eighth position, respectively.

Procedure and scoring

Each trial set was administered to a different, independent sample of participants. In all samples, participants were presented with a booklet containing instructions on the first page and ten gap-filling texts, with each text on a separate page. Texts had to be worked on in the order and within the time limits given; paging up and down the booklet was not allowed.

Each correctly restored word, or each acceptable variant (e.g., use of a plural form instead of the singular), was scored one point. Each incorrectly restored word, including spelling errors, was scored zero points (for a discussion of different scoring procedures for C-tests, see Eckes & Grotjahn, 2006b). Thus, the total score, computed across all 10 texts, could range from 0 to 200 points.

Data analysis

The RSM-based analyses were conducted using the computer program WINSTEPS (version 3.70; Linacre, 2010). The analyses based on the CRSM were conducted using the same-named computer program (version 1.3; Müller, 1999a).[2]

Aside from the particular Rasch models implemented, WINSTEPS and CRSM differ in the methods employed for estimating model parameters. Whereas WINSTEPS makes use of a joint (or unconditional) maximum likelihood (JML) method, CRSM adopts a conditional maximum likelihood (CML) approach (for a detailed discussion of parameter estimation methods, see Baker & Kim, 2004; Fischer, 2007; Linacre, 2004a, 2004b). Recent research comparing the JML and CML estimation methods in the applied context of calibrating C-test texts based on the RSM found that differences in the item parameters estimates obtained were negligibly small (Eckes, 2007).

In order to examine data-model fit, both WINSTEPS and CRSM provide users with an unweighted mean-square fit statistic (Wright & Masters, 1982). This statistic, also called *outfit* (Linacre, 2002), has an expected value of 1 and can range from 0 to infinity. Linacre (2002, 2010) suggested 0.50 as a lower-control limit and 1.50 as an upper-control limit for the outfit mean-square statistic. That is, Linacre considered mean-square values in the range between 0.50 and 1.50 as “productive for measurement” or as indicative of “useful fit” (see also Linacre, 2003). Other researchers suggested to use a narrower range defined by a lower-control limit of 0.70 (or 0.75) and an upper-control limit of 1.30 (see, e.g., Bond & Fox, 2007; R. M. Smith, 2004). The actual definition of lower- and upper-control limits for mean-square fit statistics will mainly depend on the nature of the assessment purpose (e.g., high-stakes vs. low-stakes decisions).

WINSTEPS also provides users with a weighted mean-square statistic (Wright & Masters, 1982). This statistic, also called *infit* (Linacre, 2002), has the same expected value and the same range of values as the outfit statistic. Whereas outfit is more sensitive to

[2] The CRSM program is available from the author upon request.

unexpected responses on items located away from a person's proficiency level, infit is more sensitive to unexpected responses on items near a person's proficiency level. Fit values greater than 1 (*misfit*) are generally deemed to be more problematic than fit values smaller than 1 (*overfit*), because misfit can change the substantive meaning of the resulting parameter estimates (Myford & Wolfe, 2003; Wright & Linacre, 1994).

Program CRSM does not compute an infit statistic, but yields two other fit statistics instead. The first of these is the conditional item-fit index (Rost & von Davier, 1994). This index, also called the *Q* index, ranges from 0 to 1, with smaller index values indicating higher fit. As a minimum requirement for data-model fit, *Q* should be smaller than 0.5 (Rost & von Davier, 1994).

The second statistic concerns the dispersion parameter λ. As mentioned earlier, positive values of this parameter indicate the regular case of the model; that is, $\lambda > 0$ means that the persons' responses on the items are in line with assumptions of the CRSM. Under the model, the dispersion parameter is assumed to be item-independent (i.e., $\lambda_i = \lambda$ for all items i). Yet, to enable an approximate model test at the level of each individual item, an adjusted version of the dispersion parameter can be employed. The adjusted parameter estimate $\widehat{\lambda}_i$ is computed in a stepwise fashion on the basis of λ as estimated for pairs of items.

Let (i, j) denote a pair of items i and j taken from the set of items under consideration. The program first estimates the dispersion parameter for all pairs of items separately. Then, the estimate of the adjusted dispersion parameter for item i is computed as

$$\widehat{\lambda}_i \approx \widehat{\lambda} + \frac{2}{I-2}\sum_{j \neq i}(\widehat{\lambda}_{(i,j)} - \widehat{\lambda}), \tag{5}$$

where I is the number of items in the set considered and $\widehat{\lambda}_{(i,j)}$ is the parameter estimate for item pair (i, j).

Unlike midpoint c of the rating scale, length d influences the estimates such that only the terms $\widehat{\beta}_i d$ and $\widehat{\lambda} d^2$ remain invariant. Since λd is the distance between midpoint β_i and endpoints $\beta_i \pm \lambda d$ of the hypothetical threshold distribution for item i, fixing the (additional) factor d has the effect of fixing the latent unit of measurement. When CRSM estimates are to be compared with estimates from another unidimensional polytomous Rasch model where $m + 1$ categories are scored with successive integers, as is the case in the present study, Müller (1999a) suggested to use $d = m + 1$. Accordingly, the estimates of the unadjusted parameter $\widehat{\lambda}$ as well as the estimates of the adjusted parameter $\widehat{\lambda}_i$ were multiplied by the number of response categories of the rating scale (i.e., $m + 1 = 21$).

Both WINSTEPS and the CRSM program produce summary Rasch statistics as proposed by Wright and Masters (1982): The person separation index, from which the number of person strata index *H* can be computed, and the test reliability of person separation *R*. Index *H* is of special importance when a measurement instrument is to be used for placing examinees in a number of different levels of proficiency. In the present study, *H* indicates the number of statistically distinct levels of examinee proficiency in a given sample of examinees. In general, the number of proficiency levels that a measurement instrument can reliably distinguish should be at least as high as the number of proficiency levels that the instrument purports to distinguish. Since the online placement test that is the focus of this research is intended to sort examinees into one of four levels or ordered classes of language proficiency, the *H* index should take on values of at least 4.0.

Based on the findings from the separate, sample-specific Rasch analyses, texts with suitable psychometric properties were selected for inclusion in a concurrent Rasch analysis. Generally speaking, a concurrent estimation procedure involves estimating item parameters using the data from two (or more) test forms, linked by a set of common items, simultaneously in a single run to achieve a common scale. In an alternative approach, the separate estimation procedure, item parameters are first estimated for each sample separately, and then the linear relationship of the parameter estimates for the common items is used to transform one set of parameter estimates to the scale of the other form (see, e.g., Kim & Cohen, 1998; Kolen & Brennan, 2004; Lee, Song, & Kim, 2004).

Research into the relative efficiency of these two general procedures has shown that the concurrent estimation approach often yields more precise parameter estimates (Hanson & Béguin, 2002; Lee, Song, & Kim, 2004). This finding, along with the fact that in the present research items from many test forms (i.e., 27 trial sets) had to be put on a common scale, led to the decision of using a concurrent procedure. The implementation of this procedure was accomplished by means of the RSM, using the MFORMS command available in WINSTEPS (Linacre, 2010). Program CRSM provided no such analytic facility.

Results

Descriptive sample statistics

Table 1 displays descriptive statistics for each of the 27 trial samples. The number of participants per sample ranged from 168 to 313, the average sample size was about 220. For the most part, these sample sizes can be considered sufficient for the present Rasch analysis purposes.

The mean raw scores (Column three, Table 1) ranged widely, from 63.5 points (Sample S16) to 124.3 points (Sample S06). Thus, Sample S06 participants achieved, on average, almost twice as much points as Sample S16 participants. The standard deviation of the raw score distribution varied similarly widely across samples, ranging from a minimum of 27.4 points (Sample S07) to a maximum of 47.0 points (Sample S11). Thus, although action was taken to collect data from samples with roughly similar mean proficiency level and similar within-sample proficiency variation, samples seemed to show marked differences with respect to both statistics.

Since the proficiency level of participants within a sample and the difficulty level of items within the respective trial set were confounded, definitive conclusions regarding differences in mean within-sample proficiency level could not be drawn. For example, it might have been the case that Sample S06 participants had taken the easiest trial set,

Table 1:
Descriptive statistics for 27 independent trial samples

Sample	*N*	M_o	*SD*	Max.	Min.
S01	168	114.4	34.5	198	38
S02	187	109.0	46.6	192	9
S03	276	104.4	45.1	196	8
S04	189	108.5	44.2	193	19
S05	196	121.8	40.0	197	24
S06	204	124.3	35.4	192	15
S07	168	77.6	27.4	150	6
S08	188	124.2	32.8	193	36
S09	186	108.0	37.7	190	22
S10	188	96.6	32.1	189	18
S11	222	72.3	47.0	198	4
S12	177	104.5	38.8	173	16
S13	212	104.9	35.1	182	30
S14	234	66.3	32.5	162	13
S15	230	71.9	31.7	158	10
S16	208	63.5	31.1	176	10
S17	206	104.3	33.8	175	31
S18	212	102.4	38.3	190	13
S19	205	103.6	33.6	185	16
S20	281	91.1	35.6	191	12
S21	253	81.0	33.7	172	19
S22	240	73.2	37.6	183	11
S23	212	88.2	46.1	190	8
S24	260	108.4	40.0	193	11
S25	313	102.8	32.3	186	22
S26	253	101.0	35.6	191	22
S27	259	95.3	36.1	194	10

Note. In each sample, participants worked on 10 texts with 20 gaps each (minimum score = 0, maximum score = 200). M_o = mean observed (raw) score.

whereas Sample S16 participants (possibly with much the same mean proficiency level as Sample 06 participants) had taken the most difficult one. It is only through the concurrent Rasch analysis reported later that proficiency and difficulty effects can be distinguished from one another.

In the next section, main results from the separate Rasch analyses of all 27 trial samples are summarized. Then, results from an exemplary Rasch analysis based on the RSM and the CRSM, respectively, are presented.

Separate Rasch analyses

As can be seen in Table 2, the mean person logits closely mirrored the mean raw scores shown in Table 1. For example, Samples S05, S06, and S08 had both the highest mean logits and the highest mean scores; at the opposite end of the logit and the raw score scales, respectively, were Samples S14, S15, and S16. Moreover, the mean RSM and CRSM logits were in fine agreement with each other (Pearson-$r = .97$, $p < .001$). Only in Samples S01, S13, and S15 did the absolute differences in mean logits turn out to be somewhat higher, reaching a maximum of 0.20 logits. In both Rasch-based analyses, the mean standard errors were at a fairly low level.

The number of person strata index *H* did not fall below 5.0, and in most samples ranged from 6.50 to 8.50. The high measurement precision of each of the trial sets is also indicated by the reliability index of person separation. This index ranged from .94 to .98.

Regarding the fit indices that are specific to the CRSM analyses, the estimates of the dispersion parameter clearly indicated the regular case of the model in all samples; that is, the estimated λ values did not fall at or below 0.0 in any sample. Moreover, the *Q* index computed at the level of each item in a given trial set (not shown in Table 2) ranged from .02 to .12, with the majority of *Q* values staying below .07.

Table 3 presents the frequencies of mean square fit indices (infit, outfit) for three different fit intervals, again computed at the level of individual texts within samples.

There were only two (out of the 270) cases in which the fit analysis for the RSM-estimated item parameters yielded infit and outfit statistics exceeding the upper-control limit of 1.50. In the CRSM analyses, the number of grossly misfitting texts totalled four. Concerning the narrower 0.70/1.30 interval, the vast majority of texts (i.e., over 90%) could still be considered fitting both the RSM and the CRSM. Finally, applying the 0.90/1.10 interval resulted in about 14% of the texts being diagnosed as misfitting; the number of overfitting texts increased considerably, and, in the CRSM analysis exceeded the number of fitting texts.

For purposes of illustration, Table 4 presents the RSM- and CRSM-based measurement results for a particular trial sample (i.e., Sample S18) that, as judged by the descriptive statistics listed in Table 1 and the summary Rasch statistics listed in Tables 2 and 3, could be considered typical of the set of samples studied.

Table 2:
Summary Rasch statistics for 27 independent trial samples based on Andrich's (1978) rating scale model (RSM) and Müller's (1987) continuous rating scale model (CRSM)

	RSM				CRSM				
Sample	M_l	*SE*	*H*	*R*	M_l	*SE*	*H*	*R*	$\hat{\lambda}d$
S01	0.23	0.17	6.56	.96	0.37	0.19	6.53	.96	1.98
S02	0.30	0.17	8.49	.97	0.22	0.16	8.75	.98	1.90
S03	0.10	0.17	8.69	.98	0.13	0.17	9.45	.98	2.39
S04	0.31	0.18	9.20	.98	0.26	0.17	9.79	.98	2.60
S05	0.64	0.17	7.73	.97	0.57	0.21	7.29	.96	2.00
S06	0.54	0.16	6.61	.96	0.52	0.15	7.01	.96	1.80
S07	–0.55	0.17	5.69	.94	–0.56	0.16	6.12	.95	2.33
S08	0.61	0.17	6.59	.96	0.60	0.16	6.93	.96	2.28
S09	0.22	0.16	7.31	.96	0.19	0.15	7.81	.97	2.12
S10	–0.08	0.16	6.41	.95	–0.08	0.15	6.85	.96	2.19
S11	–0.77	0.19	8.87	.98	–0.70	0.22	8.40	.97	2.21
S12	0.17	0.17	7.77	.97	0.09	0.15	8.12	.97	2.16
S13	0.31	0.16	6.95	.96	0.11	0.15	7.19	.96	1.99
S14	–0.89	0.18	6.60	.96	–0.87	0.17	7.08	.96	2.22
S15	–0.83	0.17	6.35	.95	–0.70	0.16	6.92	.96	2.31
S16	–0.90	0.18	6.40	.95	–0.89	0.17	6.70	.96	2.22
S17	0.17	0.17	6.87	.96	0.13	0.16	7.33	.96	2.18
S18	0.10	0.17	7.59	.97	0.09	0.16	7.97	.97	2.10
S19	0.12	0.17	6.93	.96	0.10	0.15	7.38	.96	2.28
S20	–0.19	0.17	7.16	.96	–0.20	0.16	7.73	.97	2.30
S21	–0.45	0.17	6.71	.96	–0.44	0.16	7.20	.96	2.19
S22	–0.67	0.18	7.76	.97	–0.72	0.17	8.68	.97	2.59
S23	–0.30	0.18	9.03	.98	–0.30	0.17	10.05	.98	2.44
S24	0.22	0.17	8.08	.97	0.22	0.16	8.71	.97	2.44
S25	0.05	0.17	6.79	.96	0.07	0.16	7.13	.96	2.39
S26	0.04	0.15	6.87	.96	0.03	0.14	7.01	.96	1.86
S27	–0.06	0.16	7.16	.96	–0.08	0.15	7.48	.97	2.06

Note. Rasch statistics M_l through *R* refer to person measures. M_l = mean logit. *SE* = mean standard error. *H* = number of person strata. *R* = test reliability of person separation. $\hat{\lambda}d$ = estimate of the dispersion parameter multiplied by length $d = m + 1$ of the rating scale (i.e., $d = 21$).

Table 3:
Frequency of infit and outfit statistics of texts from 27 trial sets using different fit intervals

	RSM				CRSM	
	Infit		Outfit		Outfit	
Interval	Freq.	%	Freq.	%	Freq.	%
Fit < 0.50	0	0.0	0	0.0	0	0.0
0.50 ≤ Fit ≤ 1.50	268	99.3	268	99.3	266	98.5
Fit > 1.50	2	0.7	2	0.7	4	1.5
Fit < 0.70	8	3.0	6	2.2	7	2.6
0.70 ≤ Fit ≤ 1.30	257	95.2	256	94.8	252	93.3
Fit > 1.30	5	1.9	8	3.0	11	4.1
Fit < 0.90	102	37.8	98	36.3	118	43.7
0.90 ≤ Fit ≤ 1.10	129	47.8	133	49.3	115	42.6
Fit > 1.10	39	14.4	39	14.4	37	13.7

Note. Each of the 27 trial sets contained 10 texts. The CRSM program produced outfit statistics only.

Table 4:
Rasch measurement results for a trial set of 10 texts (Sample S18) based on Andrich's (1978) rating scale model (RSM) and Müller's (1987) continuous rating scale model (CRSM)

Rasch statistic	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Text 7	Text 8	Text 9	Text 10
					RSM					
Measure	–1.21	–0.77	0.29	0.69	–0.15	0.01	0.19	–0.13	0.59	0.50
SE	0.04	0.04	0.03	0.04	0.03	0.03	0.03	0.03	0.04	0.04
Infit	0.85	1.18	0.92	0.93	0.76	0.90	0.96	1.10	0.95	0.91
Outfit	0.88	1.19	0.91	0.92	0.86	0.90	0.93	1.07	0.95	0.92
					CRSM					
Measure	–1.18	–0.63	0.24	0.62	–0.12	0.00	0.16	–0.11	0.53	0.49
SE	0.09	0.06	0.04	0.06	0.05	0.04	0.04	0.04	0.06	0.06
Outfit	0.93	1.17	0.90	0.88	0.85	0.88	0.89	1.03	0.92	0.85
Q index	0.05	0.07	0.04	0.04	0.05	0.04	0.04	0.05	0.06	0.05
$\hat{\lambda}_i d$	2.39	1.55	2.06	2.02	2.65	2.10	1.97	1.70	2.12	2.44

Note. Measures denote item difficulty estimates in logits. *SE* = standard error. Outfit is the (unweighted) mean-square fit statistic (Wright & Masters, 1982). The *Q* index is a fit statistic that takes on values between 0 and 1 (higher model fit is indicated by lower values; Rost & von Davier, 1994). $\hat{\lambda}_i d$ = adjusted dispersion parameter multiplied by length $d = m + 1$ of the rating scale (i.e., $d = 21$).

Text 1 and Text 2 were the easiest ones. Text 3, and particularly Text 4, proved to be more difficult for the examinees to deal with than originally expected (based on pre-testing data). As can be seen, the RSM and CRSM estimates of text difficulty were in close agreement with each other. Text 2 had a slight misfitting tendency. All other texts showed fully satisfactory data-model fit.

In order to select texts to be included in the next step of the analysis, that is, the concurrent Rasch analysis, two criteria were employed. The first selection criterion was based on the fit analysis conducted separately in each trial sample. Given that the intended placement test would be a low- to medium-stakes test, the 1.30 upper control limit was applied for both the outfit statistic (based on the RSM and the CRSM), and the infit statistic (based on the RSM only). This led to the elimination of 11 texts. The second criterion made use of the results from an analysis of differential item functioning (DIF) related to (a) examinee gender and (b) region of origin (European vs. non-European). In each and every sample, the two anchor texts proved to be of high psychometric quality (e.g., fit statistics stayed well within a narrow fit range of 0.80/1.20), thus corroborating the results of extensive pre-testing.

For purposes of the DIF analysis, the procedure implemented in WINSTEPS (Linacre, 2010) was adopted. This procedure involved first a joint run of all examinees to produce anchor values for examinee proficiency measures and for the rating scale structure (i.e., the threshold measures). Then, separate analyses were run with female and male (or European and non-European) examinee proficiency measures and the rating scale structure anchored at the values obtained in the previous analysis to produce item difficulty estimates separately for both groups of examinees. Finally, pairwise item difficulty difference *t*-tests were conducted between the two sets of item difficulty estimates (for more detail, see Linacre, 2010; see also Ferne & Rupp, 2007; Kristjansson, Aylesworth, McDowell, & Zumbo, 2005; Mapuranga, Dorans, & Middleton, 2008).

Since this method of DIF detection required 10 comparisons to be made per sample, critical significance levels had to be adjusted to guard against falsely rejecting the null hypothesis that no DIF was present. To this purpose, methods such as those based on the Bonferroni inequality (see Myers & Well, 2003) or the Benjamini–Hochberg procedure (see Thissen, Steinberg, & Kuang, 2002) can be used. Adopting the Benjamini–Hochberg approach, 12 gender-related DIF texts were identified and subsequently excluded from further analysis. Seven of these texts were significantly more difficult for females than for males, the remaining five texts were significantly more difficult for males than for females. Inspection of content did not suggest any unintended factor that could be hypothesized to account for the observed differences in item difficulty. None of the texts showed DIF related to region of origin.

Concurrent Rasch analysis

Figure 1 displays the result of the analysis in form of an examinee-text map, also called distribution map (Linacre, 2010) or Wright map (Wilson, 2005). The map illustrates that, through this analysis, all 5,927 examinees and all 195 texts selected on the basis of the

Figure 1:
Distribution map for the concurrent Rasch analysis.

```
 3             .  +
               .  |
               .  |
               .  |
              .# T|
 2            .#  +
             .##  |  14
            .###  |  16 14
           .####  |T 14 13 22 17
         .######  |  23 16 22 16 21 22
 1       .###### S+  27 14 16 17 13 07 17
       .########  |  02 13 17 27 19 25 16 20 25
     .##########  |S 11 19 18 15 21 15 19 21 15 05 11 18 23
    .###########  |  08 12 15 18 20 21 20 27 19 21 16 20 01 26 19 21 06 10 13
   .############  |  04 07 09 22 26 23 24 25 01 03 10 20 24 01 24 03 27 04 07 08 10 27 09 12
 0 .############ M+M 26 27 03 26 24 16 11 22 23 25 26 17 04 11 18 24 25 03 04 19
   .############  |  01 25 08 13 03 04 09 22 01 26 02 06 02 03 05 06 07 12 18 22 23 10 02 06 09 11 12 23 01 02 20
   .############  |  24 03 04 09 11 12 13 14 20 04 12 11 14 20 08 09 12 24 01 08
     .##########  |S 06 01 05 07 14 10 17 05 08 09 14 04 02 06 07 23 05
       .########  |  10 19 22 01 02 18 26 03
-1     .########  +  08 13 17
         .###### S|  16 21 19
          .#####  |T 05 06 07 15
           .####  |  12 17 21
             .##  |
-2           .##  +
              .# T|  06
              .#  |
               .  |
               .  |
-3             .  +
```

separate Rasch analyses were put on a common scale. This scale, the logit scale, is shown on the left-hand side. For ease of presentation, the logit scale was truncated at ± 3.0 logits.

Immediately to the right of the logit scale, the locations of the examinees are shown. These locations correspond to the estimates of the examinee proficiency measures. Each "#" indicates the location of 36 examinees, and each dot indicates the location of 1 to 35 examinees. On the right-hand side, the locations of the texts, corresponding to the text difficulty estimates, are shown. Each text is represented by a two-digit number indicating the set to which that text belonged. For example, "14" designates texts belonging to a trial set presented to Sample S14. Along the line in the middle, markers summarize the distribution of examinee and text measures, respectively. An "M" marker represents the location of the mean measure, "S" markers are placed one sample standard deviation away from the mean, and "T" markers are placed two sample standard deviations away from the mean.

Three features of the map deserve particular attention: (a) the distribution of the examinee proficiency measures lines up nicely with that of the text difficulty measures, (b) the vast majority of texts are located in the middle range of the logit scale, that is, within one sample standard deviation around the mean, and (c) the distribution of the examinee measures approximates the normal distribution. These features attest to a high potential of the text pool to differentiate between examinees in terms of the construct being measured, that is, general language proficiency.

This conclusion is corroborated by the Rasch summary statistics H and R. Considering the total sample of participants, the results were $H = 7.60$, and $R = .97$. Thus, about seven-and-a-half classes of examinees were reliably distinguished by the pool of texts studied here (i.e., almost twice as much as the final instrument is supposed to differentiate).

Finally, Table 5 presents the RSM-based calibrations of the category thresholds, which correspond to the τ_j parameter in Equation 1. The threshold measures increased in a strictly monotonic fashion from the lowest to the highest category, meaning that they were ordered as commonly required (see, e.g., Linacre, 2004b). As Luo (2005) argued, the ordering requirement is not built into the mathematical definition of the RSM, yet it ensures a proper structure of the measurement system. Thus, in the present case, filling in correctly five gaps indicated a higher proficiency than filling in correctly four gaps. Further supporting evidence of a satisfactory rating scale structure was provided by the infit and outfit statistics, defined as the mean-square fit values associated with the responses in each category. Except for the lowest two categories, these statistics were close to their expected value of 1.

Table 5:
Category threshold calibrations and fit statistics from the concurrent Rasch analysis

Category	Threshold measure	Infit	Outfit
0	–	1.23	1.10
1	–3.08	1.16	1.12
2	–2.56	1.08	1.06
3	–2.04	1.00	1.02
4	–1.58	0.99	0.99
5	–1.28	0.98	0.98
6	–1.01	0.96	0.96
7	–0.70	0.95	0.95
8	–0.49	0.94	0.93
9	–0.24	0.92	0.90
10	–0.04	0.94	0.94
11	0.22	0.93	0.93
12	0.37	0.91	0.90
13	0.59	0.92	0.93
14	0.79	0.94	0.95
15	1.09	0.92	0.93
16	1.30	0.99	1.01
17	1.57	1.01	1.02
18	1.92	0.89	0.92
19	2.24	0.97	0.98
20	2.93	0.97	0.99

Note. Threshold measures were estimated on the basis of the total sample of 5,927 participants and 195 texts.

Discussion

Summary of main findings

The present research provided strong support for the use of polytomous Rasch models for purposes of C-test item banking. In particular, Andrich's (1978) rating scale model (RSM) and Müller's (1987, 1999b) continuous rating scale model (CRSM), which extends the discrete RSM to the case of continuous responses, were successfully applied to the analysis of a series of 27 trial sets, each set containing 10 mutilated texts with 20 gaps each. The RSM was implemented by means of computer program WINSTEPS (Linacre, 2010), the CRSM by the program of the same name (Müller, 1999a). In these analyses, each text was considered a polytomous item, where the number of response categories equaled the number of gaps plus one.

Statistical indicators of data-model fit showed that, within each sample analyzed separately, the majority of texts, that is, 93% to 95% (depending on the kind of model and fit statistic) had satisfactory fit values. Moreover, parameter estimates were highly precise, as judged (a) by the number of person strata index, ranging from 5.69 to 9.20 (RSM), and from 6.12 to 10.05 (CRSM), and (b) by the reliability of person separation, ranging from .94 to .98 (RSM), and from .95 to .98 (CRSM). The adjusted dispersion parameter, a fit statistic specific to the CRSM, revealed that in each and every sample the regular case of the CRSM held. Using the DIF detection procedure available in WINSTEPS, a total of 12 texts were shown to be differentially difficult for female and male participants. No DIF effects associated with examinees' region of origin were found.

Based on the findings from the sample-specific analysis of data-model fit (fit statistics, gender-related DIF), 23 texts were excluded from further consideration. The difficulty estimates of the remaining 195 texts were put on a common scale by means of a concurrent estimation procedure in order to provide the text calibrations needed for constructing the item bank. The required link between the 27 different trial sets was provided by two texts, which were held constant across all sets.

Using the method implemented in WINSTEPS to equate multiple input files, estimates of each of the 195 text's difficulty were obtained. Results showed that the distributions of examinee proficiency measures and text difficulty calibrations were closely in line with each other. In addition, the Rasch summary statistics indicated that almost twice as much as the intended four classes of examinees could reliably be distinguished by the total set of calibrated texts. Finally, the RSM-based category threshold calibrations confirmed that the measurement system functioned properly.

Related approaches

The polytomous Rasch modeling approach advocated here is just one of a number of approaches that future research may address with respect to their utility for evaluating C-tests and constructing item banks (for an overview, see Eckes, 2010b). Thus, according to Wang and Wilson (2005), the use of polytomous test models to calibrate superitems (or testlets) may only be appropriate when the extent of LID within testlets is moderate and the test contains a large proportion of independent items. With strong LID effects present in a test, more complex approaches may be required. Wang and Wilson (2005) proposed an approach called the Rasch testlet model. This model represents each testlet by a separate dimension that is independent of the general proficiency dimension and of the other testlet-specific dimensions.

More specifically, LID within a testlet is modeled as a random effect of the interaction between persons and items. The variance of these random effects indicates the amount of the testlet effect for a given testlet; that is, the larger the testlet-specific variance, the greater the proportion of total variance of the test score distribution that is attributable to the testlet. Ideally, when the responses were perfectly in line with the local independence assumption, the estimated variance of each testlet-specific dimension would be close to zero. Note also that only the latent proficiency dimension and the random testlet effect variables are assumed to be independently normally distributed.

The Rasch testlet model assumes that the testlet-specific dimensions are independent of one another. In the Rasch subdimension model (Brandt, 2008) this assumption is given up. Hence, application of this model requires estimating the covariances between all testlet-specific dimensions. Evidence presented so far in the context of a mathematics achievement test (Brandt, 2008) suggested that the subdimension model yielded more precise parameter estimates than the Rasch testlet model.

Outside the Rasch family of C-test calibration approaches, there has been a long tradition of potentially relevant model developments, most of which represent different versions of polytomous item response models (for an excellent discussion, see Hambleton, van der Linden, & Wells, 2010; see also Ostini & Nering, 2006). Here, three more recent developments that may hold some promise for future research shall be mentioned only very briefly.

In the testlet response theory approach (Bradlow, Wainer, & X. Wang, 1999; see also Wainer, Bradlow, & X. Wang, 2007), items are treated as random effects, with item difficulty and item discrimination parameters (in the two-parameter logistic model) following a normal distribution. However, in many applications these distributional assumptions appear to be overly restrictive.

Ferrando (2001) proposed a unidimensional item response model that extends the linear congeneric model to a nonlinear model. Similar to Müller's (1987) CRSM, the nonlinear model takes the bounded nature of the response variable into account and assumes a truncated distribution which arises from a redistribution of the latent responses that are beyond the endpoints of the rating scale.

Noel and Dauvier (2007) assumed an interpolation response mechanism according to which respondents put a check at a distance to the left boundary that is proportional to the relative value they give to the extreme answer at the opposite end of the scale. From this assumption the authors derived a one-parameter and a two-parameter beta distributional model for the manifest response variable.

Conclusion

C-tests have gained great popularity among researchers, language test providers, and language teachers alike. Based on findings from the present and related research, a completely web-based placement test of German as a foreign language has been developed showing the following key features: access to a calibrated item bank containing a large structured set of C-test texts, automatic test administration and scoring, and immediate feedback of test results. Since its launch in October 2006, this test, which (in German) is called the “Online-Einstufungstest Deutsch als Fremdsprache” (onDaF, for short; see Eckes, 2010a; see also www.ondaf.de), has been used worldwide to assign L2 learners of German to language courses at institutions of higher education, to provide feedback to L2 learners who plan to take the TestDaF (Test of German as a Foreign Language; see www.testdaf.de), and to assist lecturers in deciding on foreign students’ eligibility for scholarships of the German Academic Exchange Service (DAAD). The polytomous Rasch modeling approach to the calibration of C-test texts has formed the centerpiece of this innovative test development.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Baghaei, P. (2008). The effects of the rhetorical organization of texts on the C-test construct: A Rasch modelling study. *Melbourne Papers in Language and Testing*, *13*, 32-51.

Baghaei, P. (2010). A comparison of three polychotomous Rasch models for super-item analysis. *Psychological Test and Assessment Modeling*, *52*, 313-322.

Baghaei, P. (2011). *C-test construct validation: A Rasch modeling approach*. Saarbrücken, Germany: VDM.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Dekker.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *1*, 51-69.

Cureton, E. E. (1965). Reliability and validity: Basic assumptions and experimental design. *Educational and Psychological Measurement*, *25*, 326-346.

Eckes, T. (2006). Rasch-Modelle zur C-Test-Skalierung [Rasch models for scaling of C-tests]. In R. Grotjahn (Ed.), *Der C-Test: Theorie, Empirie, Anwendungen/The C-test: Theory, empirical research, applications* (pp. 1-44). Frankfurt, Germany: Lang.

Eckes, T. (2007). Konstruktion und Analyse von C-Tests mit Ratingskalen-Rasch-Modellen [Construction and analysis of C-tests with rating scale Rasch models]. *Diagnostica*, *53*, 68-82.

Eckes, T. (2010a). Der Online-Einstufungstest Deutsch als Fremdsprache (onDaF): Theoretische Grundlagen, Konstruktion und Validierung [The online placement test of German as a foreign language (onDaF): Theoretical foundations, construction, and validation]. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-test: Contributions from current research* (pp. 125-192). Frankfurt, Germany: Lang.

Eckes, T. (2010b). Rasch models for C-tests: Closing the gap on modern psychometric theory. In A. Berndt & K. Kleppin (Eds.), *Sprachlehrforschung: Theorie und Empirie – Festschrift für Rüdiger Grotjahn* (pp. 39-49). Frankfurt, Germany: Lang.

Eckes, T., & Grotjahn, R. (2006a). A closer look at the construct validity of C-tests. *Language Testing*, *23*, 290-325.

Eckes, T., & Grotjahn, R. (2006b). C-Tests als Anker für TestDaF: Rasch-Analysen mit dem kontinuierlichen Ratingskalen-Modell [C-tests as anchor for TestDaF: Rasch analyses using the continuous rating scale model]. In R. Grotjahn (Ed.), *Der C-Test: Theorie, Empirie, Anwendungen/The C-test: Theory, empirical research, applications* (pp. 167-193). Frankfurt, Germany: Lang.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4*, 113-148.

Ferrando, P. J. (2001). A nonlinear congeneric model for continuous item responses. *British Journal of Mathematical and Statistical Psychology*, *54*, 293-313.

Fischer, G. H. (2007). Rasch models. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (*Handbook of statistics*, Vol. 26, pp. 515-585). Amsterdam: Elsevier.

Grotjahn, R. (1987). How to construct and evaluate a C-test: A discussion of some problems and some statistical analyses. In R. Grotjahn, C. Klein-Braley & D. K. Stevenson (Eds.), *Taking their measure: The validity and validation of language tests* (pp. 219-253). Bochum: Brockmeyer.

Grotjahn, R., Klein-Braley, C., & Raatz, U. (2002). C-tests: An overview. In J. A. Coleman, R. Grotjahn & U. Raatz (Eds.), *University language testing and the C-test* (pp. 93-114). Bochum: AKS-Verlag.

Hambleton, R. K., van der Linden, W. J., & Wells, C. S. (2010). IRT models for the analysis of polytomously scored data: Brief and selected history of model building advances. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 21-42). New York: Routledge.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*, 3-24.

Harsch, C., & Schröder, K. (2007). Textrekonstruktion: C-Test [Text reconstruction: C-test]. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen: Konzepte und Messung* (pp. 212-225). Weinheim, Germany: Beltz.

Henning, G. (1989). Meanings and implications of the principle of local independence. *Language Testing*, *6*, 95-108.

Kim, S.-H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*, 131-143.

Klein-Braley, C. (1997). C-tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*, 47-84.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, *65*, 935-953.

Lee, W.-C., Song, M.-Y., & Kim, J.-P. (2004, April). *An investigation of procedures for obtaining a common IRT scale*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Lee-Ellis, S. (2009). The development and validation of a Korean C-test using Rasch analysis. *Language Testing*, *26*, 245-274.

Linacre, J. M. (2001). Percentages with continuous Rasch models. *Rasch Measurement Transactions*, *14*, 771-774.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.

Linacre, J. M. (2003). Size vs. significance: Infit and outfit mean-square and standardized chi-square fit statistic. *Rasch Measurement Transactions*, *17*, 918.

Linacre, J. M. (2004a). Estimation methods for Rasch measures. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 25-47). Maple Grove, MN: JAM Press.

Linacre, J. M. (2004b). Optimizing rating scale category effectiveness. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258-278). Maple Grove, MN: JAM Press.

Linacre, J. M. (2004c). Rasch model estimation: Further topics. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 48-72). Maple Grove, MN: JAM Press.

Linacre, J. M. (2010). *A user's guide to WINSTEPS: Rasch-model computer programs* [Computer software manual]. Chicago: Winsteps.com.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, *4*, 73-79.

Luo, G. (2005). The relationship between the rating scale and partial credit models and the implication of disordered thresholds of the Rasch models for polytomous responses. *Journal of Applied Measurement*, *6*, 443-455.

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). *A review of recent developments in differential item functioning* (Research Report, RR-08-43). Princeton, NJ: Educational Testing Service.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika*, *52*, 165-181.

Müller, H. (1999a). CRSM: A Fortran program for the analysis of continuous rating scale data according to a Rasch model for continuous responses (Version 1.3) [Computer software]. Department of Psychology, University of Vienna, Vienna, Austria.

Müller, H. (1999b). *Probabilistische Testmodelle für diskrete und kontinuierliche Ratingskalen: Einführung in die Item-Response-Theorie für abgestufte und kontinuierliche Items* [Probabilistic test models for discrete and continuous rating scales: An introduction to item response theory for graded and continuous items]. Bern: Huber.

Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis* (2nd ed.). Mahwah, NJ: Erlbaum.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*, 386-422.

Noel, Y., & Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement*, *31*, 47-73.

Norris, J. M. (2006). Development and evaluation of a curriculum-based German C-test for placement purposes. In R. Grotjahn (Ed.), *Der C-Test: Theorie, Empirie, Anwendungen/The C-test: Theory, empirical research, applications* (pp. 45-83). Frankfurt, Germany: Lang.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1-15.

Raatz, U. (1985). Better theory for better tests? *Language Testing*, *2*, 60-75.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press. (Original work published 1960)

Reichert, M., Keller, U., & Martin, R. (2009). The C-test, the TCF and the CEFR: A validation study. In R. Grotjahn (Hrsg.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-test: Contributions from current research*. Frankfurt, Germany: Lang.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika*, *53*, 349-359.

Rost, J., & von Davier, M. (1994). A conditional item-fit index for Rasch models. *Applied Psychological Measurement*, *18*, 171-182.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, *38*, 203-219.

Sigott, G. (2004). *Towards identifying the C-test construct*. Frankfurt, Germany: Lang.

Smith, E. V. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement*, *6*, 147-163.

Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73-92). Maple Grove, MN: JAM Press.

Spolsky, B. (1971). Reduced redundancy as a language testing tool. In G. E. Perren & J. L. M. Trim (Eds.), *Applications of linguistics* (pp. 383-390). Cambridge, UK: Cambridge University Press.

Szabó, G. (2008). *Applying item response theory in language test item bank building*. Frankfurt, Germany: Lang.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, *27*, 77-83.

Vale, C. D. (2006). Computerized item banking. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 261-285). Mahwah, NJ: Erlbaum.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185-201.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 126-149.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, *60*, 181-198.

Wolfe, E. W. (2000). Equating and item banking with the Rasch model. *Journal of Applied Measurement*, *1*, 409-434.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Westport, CT: American Council on Education/Praeger.

Appendix

Sample Text 1

Fragen zur Berufswahl

Alte Berufe verschwinden, neue kommen hinzu: Bei d________ Berufswahl ha________ Schüler im________ wieder Fra________ oder Prob________. Denn e________ gibt ei________ große Anz________ sehr versch________ Berufe, u________ es i________ nicht ein________, die rich________ Wahl z________ treffen. D________ berufliche Zuk________ sollte m________ rechtzeitig pla________. Dabei ka________ es sinn________ sein, sich beim Arbeitsamt beraten zu lassen. Manchmal hilft auch ein Test zu den persönlichen Berufsinteressen.

Sample Text 2

Geschichte der Familie

Familien haben ihre eigene Geschichte. Oft ka________ man s________ bis z________ einem se________ frühen Zeitp________ zurückverfolgen. D________ Älteren erzä________ gerne v________ ihrer Kind________ und Jug________. Alte Fot________ und Bri________ enthalten wich________ Informationen üb________ die Vergang________, alte Werkz________ und Masch________ zeigen, w________ Vorfahren gel________ und gearb________ haben. So kann man auf unterhaltsame Weise viel über die Geschichte der eigenen Familie erfahren.